DOCUMENT RESUME

ED 024 414 LI 001 122

By- Lufkin, Richard C.

Determination and Analysis of Some Parameters Affecting the Subject Indexing Process.

Massachusetts Inst. of Tech., Cambridge. Electronic Systems Lab.
Spons Agency- Department of Defense, Washington, D.C. Advanced Research Projects Agency.; National Science Foundation, Washington, D.C.
Report No- ESL-R-364
Pub Date Sep 68
Grant- MIT-DSR-P-70054; NSFC-472 (Part)
Note- 55p.
EDRS Price MF-$0.25 HC-$2.85
Descriptors- Cataloging, Classification, *Indexing, *Performance Factors, *Statistical Analysis, *Task Analysis
Identifiers- *Information Transfer Experiment, Intrex

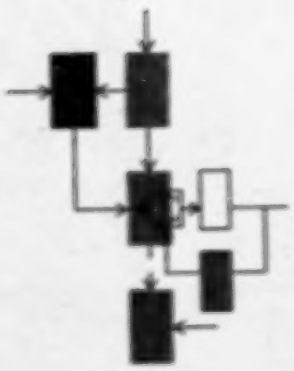
This report describes a quantitative study of subject-indexing as it is being performed by catalogers who are developing an augmented catalog as part of Project Intrex. Learning curves for the average time taken to index technical documents on a per-page basis were derived for a group of indexers. The average indexing times per page were also calculated and compared under several criteria pertaining to the nature of the documents. (Author)

EDO 24414

LI 001122
Copy 1
Dr. Barchund
NOV 20 1968

September, 1968

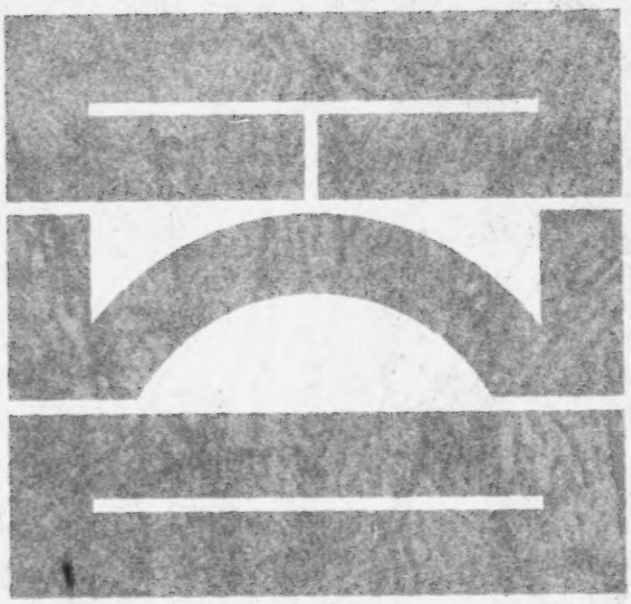
REPORT ESL-R-364
M.I.T. PROJECT DSR 70054
Research Grant NSFC-472 (Part)



DETERMINATION AND ANALYSIS OF SOME PARAMETERS AFFECTING THE SUBJECT INDEXING PROCESS

ERIC/CLIS
DEC 10 '68
UNIVERSITY OF MINNESOTA

Richard C. Lufkin



Project Intrex Group

Electronic Systems Laboratory

MASSACHUSETTS INSTITUTE OF TECHNOLOGY, CAMBRIDGE, MASSACHUSETTS 02139

Department of Electrical Engineering

LI 001122

September, 1968 Report ESL-R-364

DETERMINATION AND ANALYSIS OF SOME PARAMETERS AFFECTING THE SUBJECT INDEXING PROCESS

by

Richard C. Lufkin

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

The work described in this document was performed as part of Project Intrex under Research Grant NSFC-472 (Part) awarded to the Massachusetts Institute of Technology by the National Science Foundation and the Advanced Research Projects Agency of the Department of Defense. This grant is designated as M.I.T. DSR Project No. 70054.

Electronic Systems Laboratory
Department of Electrical Engineering
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

LI 001 122

ERRATA SHEET

Electronic Systems Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

for

Report ESL-R-364
September, 1968

"Determination and Analysis of
Some Parameters Affecting the Subject Indexing Process"

by

Richard C. Lufkin

Page 19, Figure 9:

Change the label of the horizontal axis of Fig. 9

from: "Months of Intrex Experience"

to: "Document Length, Pages"

UNIVERSITY OF MINNESOTA
JAN 23 '69
ERIC/CLIS

FOREWORD

Except for editorial changes, this report is the thesis submitted by Mr. Richard C. Lufkin to the Electrical Engineering Department, Massachusetts Institute of Technology, in partial fulfillment of the requirements for the degree of Bachelor of Science. A few alterations in the original wording have been made throughout the text in an effort to enhance clarity, and several pages have been reformatted; otherwise the manuscript remains as submitted.

J. F. Reintjes
Professor of Electrical Engineering

ABSTRACT

This report describes a quantitative study of subject-indexing as it is being performed by catalogers who are developing an augmented catalog as part of Project Intrex. Learning curves for the average time taken to index technical documents on a per-page basis were derived for a group of indexers. The average indexing times per page were also calculated and compared under several criteria pertaining to the nature of the documents.

ACKNOWLEDGMENT

I wish to thank Mr. Alan Benenfeld for his patience and time as he directed me into the rigors of documentation.

I wish to express my appreciation to Professor J. Francis Reintjes for his incisive commentaries and the demands for excellence which he imposes.

I also wish to thank Miss Leslie Smith for her assistance in punched-card preparation, Mr. Francis Heile for his help in computer-program preparation, and Mr. Steven Morrow for his assistance and steady hand in graphical presentations.

CONTENTS

		page
I	INTRODUCTION	1
II	SUBJECT INDEXING	2
III	PARAMETERS STUDIED	3
IV	DATA ACQUISITION AND PROCESSING	5
V	RESULTS	8
VI	CONCLUSION	11
	BIBLIOGRAPHY	12
	APPENDICES:	
	A. Figures and Data Tabulations	13
	B. Summary of Personal Data by Code Number	41
	C. Sample Recording Form	43
	D. Listing of Files from which Data were Gathered	44
	E. Computer Program	45
	F. A Note on Statistics	46

LIST OF FIGURES

1. Composite Learning Curves of the Catalogers page 14

2. Time History of the Average Number of Subject Terms Assigned per page 14

3. Learning Curve and Other Time History Curves for Indexer 1 15

4. Learning Curve and Other Time History Curves for Indexer 2 15

5. Learning Curve and Other Time History Curves for Indexer 7 16

6. Learning Curve and Other Time History Curves for Indexer 9 17

7. Learning Curve and Other Time History Curves for Indexer 10 17

8. Learning Curve and Other Time History Curves for Indexer 13 18

9. Subject Indexing Time, Review Time, and Number of Subject Terms, per page, versus Document Length 19

LIST OF TABLES

1.	Learning Curve Data for All Indexers	page 20
2.	Learning Curve Data for All Student Indexers	20
3.	Learning Curve Data for All Librarian Indexers	21
4.	Time History Data of the Mean per page Number of Subject Terms Assigned by All Indexers	22
5.	Learning Curve Data of Individual Indexers	23
6.	Time History Data of the Mean Review Time for Individual Indexers	27
7.	Time History Data of the Mean per page Number of Subject Terms Assigned by Individual Indexers	31
8.	Average Indexing Time, per page, versus Document Length, for All Experienced Indexers	35
9.	Average Review Time, per page, versus Document Length, for All Experienced Indexers	36
10.	Average per page Number of Subject Terms Assigned versus Document Length, for All Experienced Indexers	37
11.	Average per page Indexing Time of Documents Grouped by Format	38
12.	Average per page Number of Assigned Subject Terms for Documents Grouped by Format	38
13.	Average per page Indexing Time of Documents Grouped by the Author's Level of Approach	39
14.	Average per page Indexing Time of Documents Grouped by Subject Area	39
15.	Average per page Indexing Time of Documents Grouped by Author's Purpose	40

I. INTRODUCTION

The objective of this research is to study the effects of (1) indexer experience over time and (2) document-related characteristics on the subject-indexing process for the data base of the Project Intrex augmented catalog. A major feature of this experimental catalog is in-depth subject indexing of documents. This indexing is the most time-consuming and intellectually demanding task in the creation of catalog records for documents selected for the Intrex data base. The thought processes involved in indexing have yet to be understood. Consequently, qualitative and quantitative measurements of the intellectual indexing task are very difficult. This report studies the way the time to index individual documents for the Intrex data base varies over time. The assumption that indexers exhibit a learning period is tested. In particular, learning curves for the individual indexers are derived to show the cumulative effect of indexing experience on document indexing time. The assumption that document-related parameters have an effect on the indexing time also is studied.

II. SUBJECT INDEXING

The Intrex cataloging function is divided into two areas, descriptive cataloging and subject cataloging. Subject indexing forms the major part of subject cataloging in terms of time consumed.

The subject indexing of a document consists of the creation of a set of subject terms, each generally a combination of phrases, which represent the subject concepts discussed in the document. Terms are primarily based upon the text of a document, that is, upon the author's own words. These text-based expressions may be restructured by the indexer to provide an intensive and complete representation of a concept. A relevance weight is attached to each term as a means of indicating the extent to which the concept represented by the term is discussed in the document.

There are five steps in the indexing process:

1. Scanning the document to get an idea of its coverage
2. Identifying those concepts in the document which are to be indexed
3. Formulating an appropriately structured subject term to represent the concept (words and phrases not appearing in a document may be used)
4. Weighting the subject term
5. Recording the final decision

Ideally, in order to achieve consistency in concept selection, term structuring, and term weighting, the indexing operation should be performed identically by all indexers. In an effort to achieve this ideal, a review of the initial indexing is performed by another indexer. The reviewer checks the initial indexing for overall coverage, completeness, redundancy, accuracy, and the like, and makes corrections.

Indexing is a technique learned only through experience. New indexers go through a training period during which their work is reviewed intensively until a satisfactory level of performance is reached. With part-time student indexers, the procedure is to assign each student to a professional librarian who provides on-the-job guidance for the student. He also receives a few brief formal training sessions.

III. PARAMETERS STUDIED

The parameters of interest fall into two classes: those pertaining to the indexer and those pertaining to the nature of the document.

Parameters pertaining to the individual indexers as bases of measurement and comparison are: status as a librarian or student, amount of indexing experience prior to employment with Intrex, highest degree level attained or highest school level, and major field of college study. On the basis of actual data, all criteria except the first were later rejected as parameters of study because there was little or no variation in these parameters among the small number of indexers employed.

Measurements made with status as a criterion were directed toward the creation of a learning curve. The learning curve is defined as the time to index one page of a document as a function of the time span of experience. The dependent variable is the indexing time and this was normalized over the number of pages in a given document to eliminate the effects of document length on indexing time. The independent variable, time span of indexing experience, or the time since beginning employment by Project Intrex, was considered in one-month blocks. The latter procedure eased data gathering and allowed a more meaningful data analysis because of the availability of larger document populations per time period. Use of a smaller time span can influence the average indexing time because of the appearance of clusters of the same type of document during a short indexing period.

Additionally, the number of subject terms assigned per page per document and the per page review time per document were also measured as a function of the time span of indexing experience.

The learning curves, subject-term curves, and review-time curves were measured for each individual indexer and these data were then grouped to form composite curves for all librarians, all students, and all indexers combined.

Parameters considered pertinent to the nature of the document were:

1. Number of pages
2. Document format (arrangement of information within the document)
3. Author's purpose in writing the document
4. Level of approach (academic level of the author's intended audience)
5. Subject area for which the document was selected

Language of the document was considered but rejected as a parameter due to the very small number of foreign-language documents in the data base.

The above criteria were selected to test the assumption that document-related parameters will significantly affect the indexing time per document page. The measurements which were made for each document-related parameter are shown below.

Measurements Made for Each Document-Related Parameter

Measurement \ Parameter	Number of Pages	Subject Area	Format	Author's Purpose	Level of Approach
Indexing time per page	X	X	X	X	X
Review time per page	X				
Number of subject terms per page	X		X		

Because this phase of the study was performed after establishing the existence of learning curves, these variables were measured only for those documents indexed by librarians after three months of employment and by students after two months of employment. This plan eliminated the initial learning experiences of indexers.

IV. DATA ACQUISITION AND PROCESSING

Personal data describing the indexer were obtained through interviews. Specific information gathered was:

1. Name
2. Intrex code number
3. Professional status (librarian or student)
4. Month of initial employment by Project Intrex as an indexer
5. Degree status
6. Educational background (field of college study)
7. Amount of indexing experience prior to employment with Intrex

These data are summarized in Appendix B under the indexer's Intrex code number.

Document-related data were retrieved manually from computer printouts of catalog records. Computer search-and-sort routines were considered as a means of gathering the data, but this method was rejected because of programming inexperience and cost of computer time. Data were transferred to a specially developed recording form (see Appendix C, with sample data). The total set of data acquired from a catalog record is listed below in the order in which it is recorded on the form shown in Appendix C.

1. Record number -- this is an assigned document control number
2. Subject area -- a code number signifying the database subject area for which the document was selected
3. Number of pages
4. Field 31 code -- a letter code specifying the document format
5. Field 65 code -- a letter code specifying the author's purpose in writing the document
6. Field 66 code -- a number code specifying the author's level of approach (the academic level of the intended reader audience)
7. Number of index terms in field 73

8. Indexer -- the Intrex code number for the person who created the set of subject index terms
9. Date indexed
10. Month number -- number of the month in which the document was indexed, month one being the indexer's initial month of employment
11. Subject indexing time -- number of minutes spent subject cataloging
12. Special -- a code number used to designate documents not describable within the recording limitations of the form
13. Reviewer -- the Intrex code number for the person who reviewed the initial indexing
14. Review time -- number of minutes taken to review indexing

Under item 11 the only data recorded in the permanent cataloging file are the time spent on subject cataloging. Subject-indexing time represents the majority of this time. Time to accomplish other subject cataloging tasks was assumed to be constant for all documents and short (on the order of one or two minutes for a document). Therefore, in this report subject-cataloging time is considered to be subject-indexing time.

Data from approximately 2,500 catalog records[1] were recorded on the special form. The indexing for these records had been performed in the period March 1967 to February 1968. The only criterion for excluding a catalog record from this study was the absence of indexing terms on a record; this would be the case, for example, for records describing an entire journal title. Data were transferred to punched-card format on a standard eighty-column card to facilitate data sorting and processing. All data referred to in the above list, with the exception of item 9, were punched. Card-column assignments are noted at the bottom of the recording form shown in Appendix C.

A short computer program, written in the Fortran IV language, compiled the statistics of interest. This program is listed in Appendix E. Using indexing time per document page, review time per document page, or the number of subject terms per document page, as the

---

[1] See Appendix D for a listing of files from which data were selected. A file is a group of ten catalog records.

relevant random variable, the program calculates the average and the standard deviation for the variable for all cards in groups defined by delimiting cards.[2] The program also calculates the number of records and the average number of pages for the documents represented in the data groupings. Data sorting into groupings was done mechanically under the various classifications of data parameters, as for example, a composite of all indexers by months or the categories representing the different possible types of document format.

---

[2] A description of the statistical theory utilized in this analysis is given in Appendix F.

V. RESULTS

Figure 1* shows the composite learning curves which were derived from the study. There is a definite trend here in that, for the librarians, an initial learning period seems to end after three months; for the students the learning period ends after two months. The wide spread of mean values over time justifies our initial statement that a quantitative measure of the indexing process is very difficult. The data for the points plotted on the composite learning curves are listed in Tables 1, 2, and 3. From these tables, a trend of decreasing standard deviation over time can be seen to parallel the indicated learning curve trend. These trends indicate the time development of increasing consistency in average indexing time per page.

Figure 2, which is a plot of the data in Table 4 of the average number of subject terms chosen to describe a document as a function of months of experience, shows little or no change in the number of subject terms selected over time. When this information is combined with the Fig. 1 data, we note that while the indexing time per page decreases, the number of terms generated per page remains constant; thus the number of terms created per minute increases. Since the number of indexers whose work is represented by a given month's data grows smaller with an increase in month number (staff buildup was gradual), data towards the higher experience numbers tend more to reflect the indexing characteristics of the two or three indexers employed for the longest time.

Data for the individual indexers give an interesting insight into the average individual indexing performances per document per month. These data, for the librarians only, are graphed in Figs. 3 through 8; corresponding data for all indexers are listed in Tables 5 through 7. One should avoid making too many generalization from these data since the standard deviation is of the same order of magnitude as the

---

*All figures and tables are located in Appendix A.

mean. In addition, careful analysis of the data should take cognizance of the number of documents represented by any data point in order to judge its significance. The graphs show, with few exceptions for any one individual, that the indexing time per page, the number of subject terms assigned per page, and the review time per page, follow the same trends over time. Index time per page and review time per page are derived from the similar processes of indexing and review; both of these time averages per page reflect the number of subject terms selected per page. These variables could also be affected by sequences of difficult-to-index documents.

Comparisons of individual learning curve data show variations in the learning experience for all indexers. The indexers represented in Figs. 3 and 4, where initial pronounced learning curves are absent, had the benefit of working on the initial pilot efforts in developing the indexing method. Although there is no initial learning curve maximum in Fig. 6, the presence of an initial learning period is still suggested from the upward swing of the subject term curve from initially low values and from the decrease of the initially pronounced review time per page.

Data pertaining to the document-related parameters is presented in Fig. 9 and in Tables 8 through 15. In Fig. 9, indexing time per page, review time per page, and number of subject terms per page, all show a consistent decrease with increasing number of document pages. The standard deviations for these variables also show a decreasing trend. The large change in the times for a one- and two-page document reflects two factors which are nearly independent of document length: the time to complete the other cataloging tasks associated with subject cataloging time, and the time for the indexers' initial familiarization with the document.

The approximate ratio of review time per page to indexing time per page, as a function of document length, is 0.7 for one-page documents, 0.5-0.6 for two- and three-page documents, and 0.3-0.4 for documents of four or more pages.

Data reduction under the classification by format, shown in Table 11, confirms the assumption that document related parameters affect the time to index. For example, the percentage differential of

indexing time per page between a letters-journal article and a standard journal article is about 27 percent, whereas the differential should be about 14 percent if one considers only the different average-number-of-pages criterion. Thus, letter-type articles are more difficult to index than regular-type articles. Similarly, conference-proceedings articles are much easier to index than letters-journal articles, about 45 percent less time being spent per page; in this case, the two types of articles have about the same average number of pages, and using this as a criterion, there should be no differential in indexing time. Conference proceedings articles seem to be the easiest to index among the three major categories, since in comparison with articles from a standard journal there is a 25 percent shorter time to index per page, whereas an 11 percent greater time is predicted on a number-of-pages basis. Similar conclusions apply to Table 12, which shows the average number of subject terms per page for documents classified again by format.

VI. CONCLUSION

This analysis verifies the assumption that the subject-indexing process exhibits a learning period. This period seems to span three months for librarians and two months for students. During this initial learning period, the students also have a lower average time to index documents on a per-page basis. Indexing time appears to level off at six to eight minutes per page for an experienced worker. The rather large spread in the data emphasizes the very subjective nature of the work. Indexers who show the most consistency in indexing time over their months of association with Intrex also have the lowest average values of per page indexing time.

Analysis of data pertaining to the document-related parameters tells us that indexing time, review time, and number of subject terms, per page, are all dependent on and become more consistent with increasing document length. There are also observable differentials of indexing time per page among documents falling into different format categories. These differentials give us some insight into the effects of different types of documents on the subject indexing process.

BIBLIOGRAPHY

1. Benenfeld, A. R., Generation and Encoding of the Project Intrex Augmented Catalog Data Base, Electronic Systems Laboratory Report ESL-R-360, Massachusetts Institute of Technology, Cambridge, Mass., August, 1968, and also a paper presented at the 6th Annual Clinic on Library Applications of Data Processing, University of Illinois, May 7, 1968.

2. Benenfeld, A. R., Gurley, E. J., and Rust, J. E., Cataloging Manual, Electronic Systems Laboratory Technical Memorandum ESL-TM-303, Massachusetts Institute of Technology, Cambridge, Mass., February, 1967.

3. Overhage, Carl F. J., and Harmon, R. J., (eds.) Intrex—Report of a Planning Conference on Information Transfer Experiments, M.I.T. Press, Cambridge, Mass., 1965.

4. Project Intrex, Semiannual Activity Reports (PR-4, 15 September 1967; and PR-5, 15 March 1968) Massachusetts Institute of Technology, Cambridge, Mass.

5. Stuntz, S. C., Inputting Process to an Augmented Library Catalog, an Efficiency Study, B.S. Thesis, Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, Mass., June, 1967.

APPENDIX A

Figures and Data Tabulations

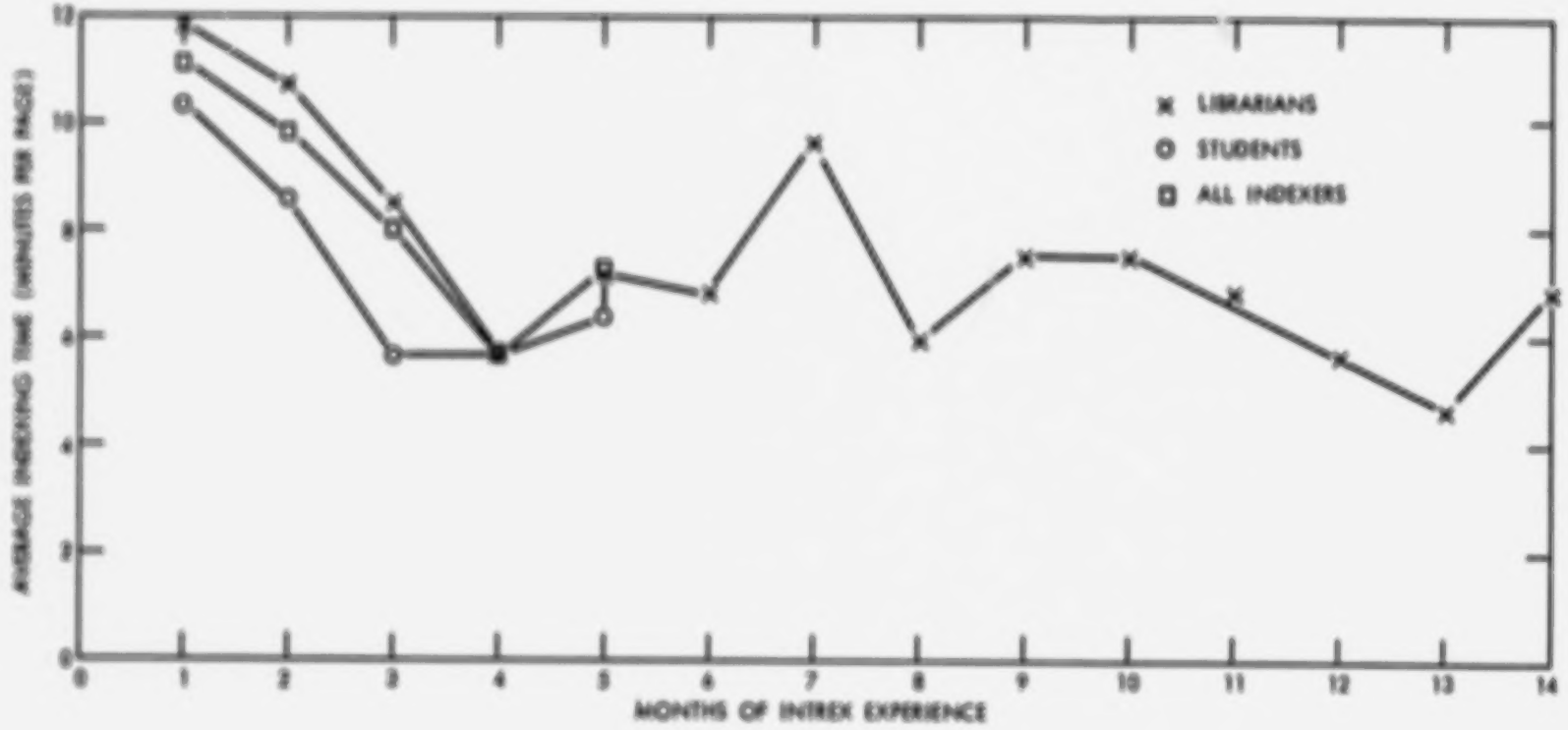


Fig. 1 Composite Learning Curves of the Catalogers

Fig. 2 Time History of the Average Number of Subject Terms Assigned per page

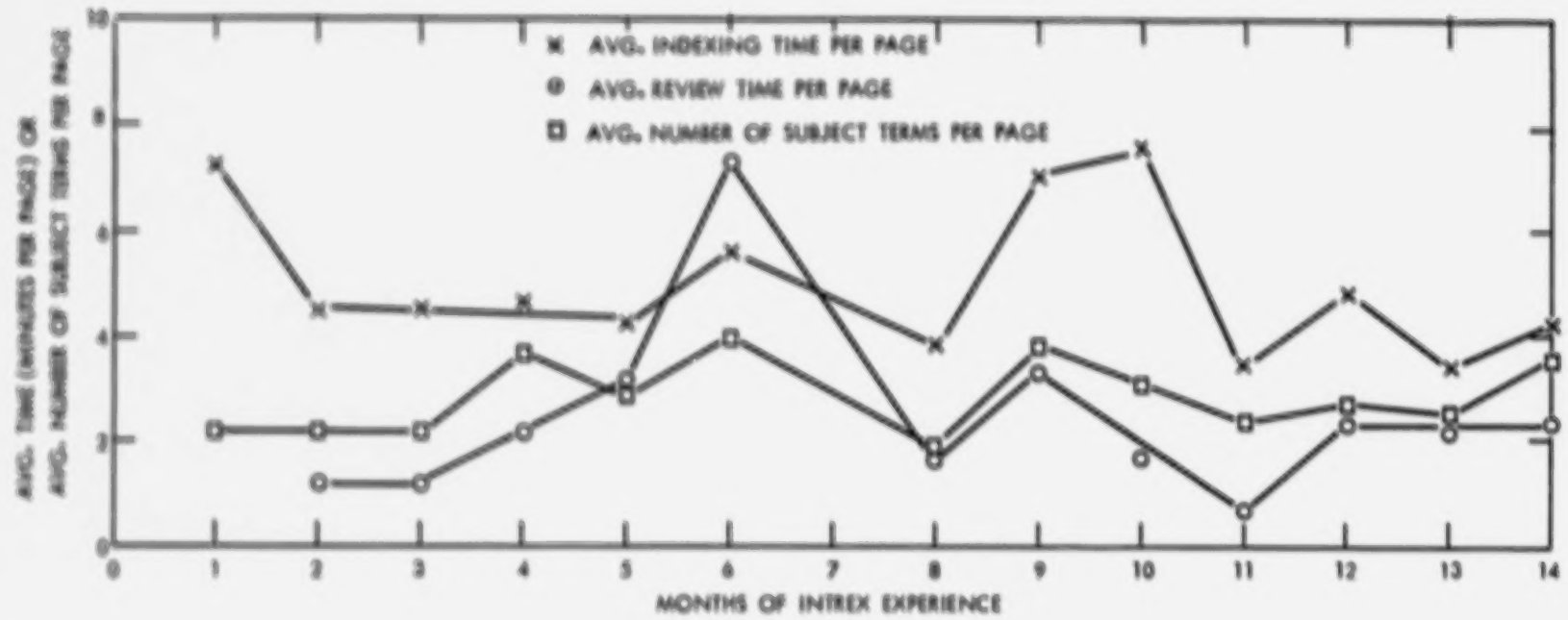

AVG. INDEXING TIME PER PAGE
AVG. REVIEW TIME PER PAGE
AVG. NUMBER OF SUBJECT TERMS PER PAGE
AVG. TIME (MINUTES PER PAGE) OR
AVG. NUMBER OF SUBJECT TERMS PER PAGE
0
2
4
6
8
10
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
MONTHS OF INTREX EXPERIENCE

Fig. 3 Learning Curve and Other Time History Curves for Indexer 1

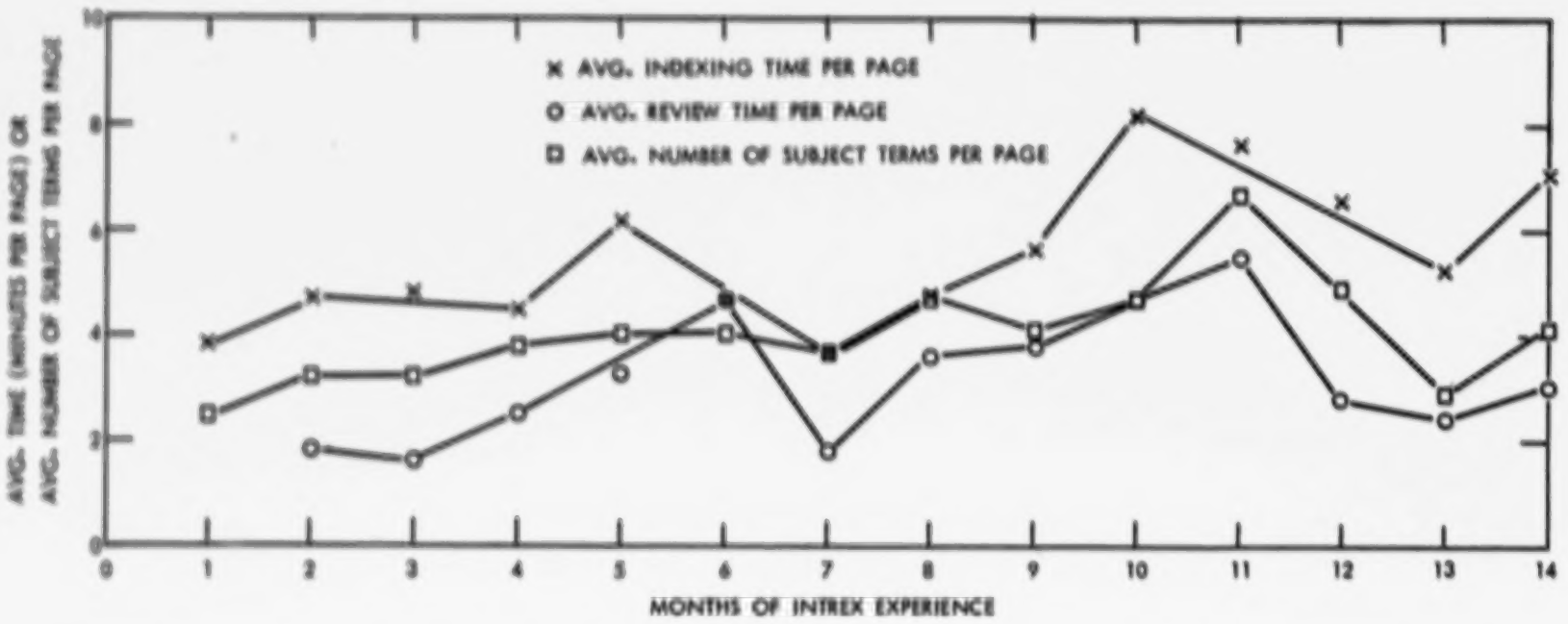

AVG. INDEXING TIME PER PAGE
AVG. REVIEW TIME PER PAGE
AVG. NUMBER OF SUBJECT TERMS PER PAGE
AVG. TIME (MINUTES PER PAGE) OR
AVG. NUMBER OF SUBJECT TERMS PER PAGE
0
2
4
6
8
10
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
MONTHS OF INTREX EXPERIENCE

Fig. 4 Learning Curve and Other Time History Curves for Indexer 2

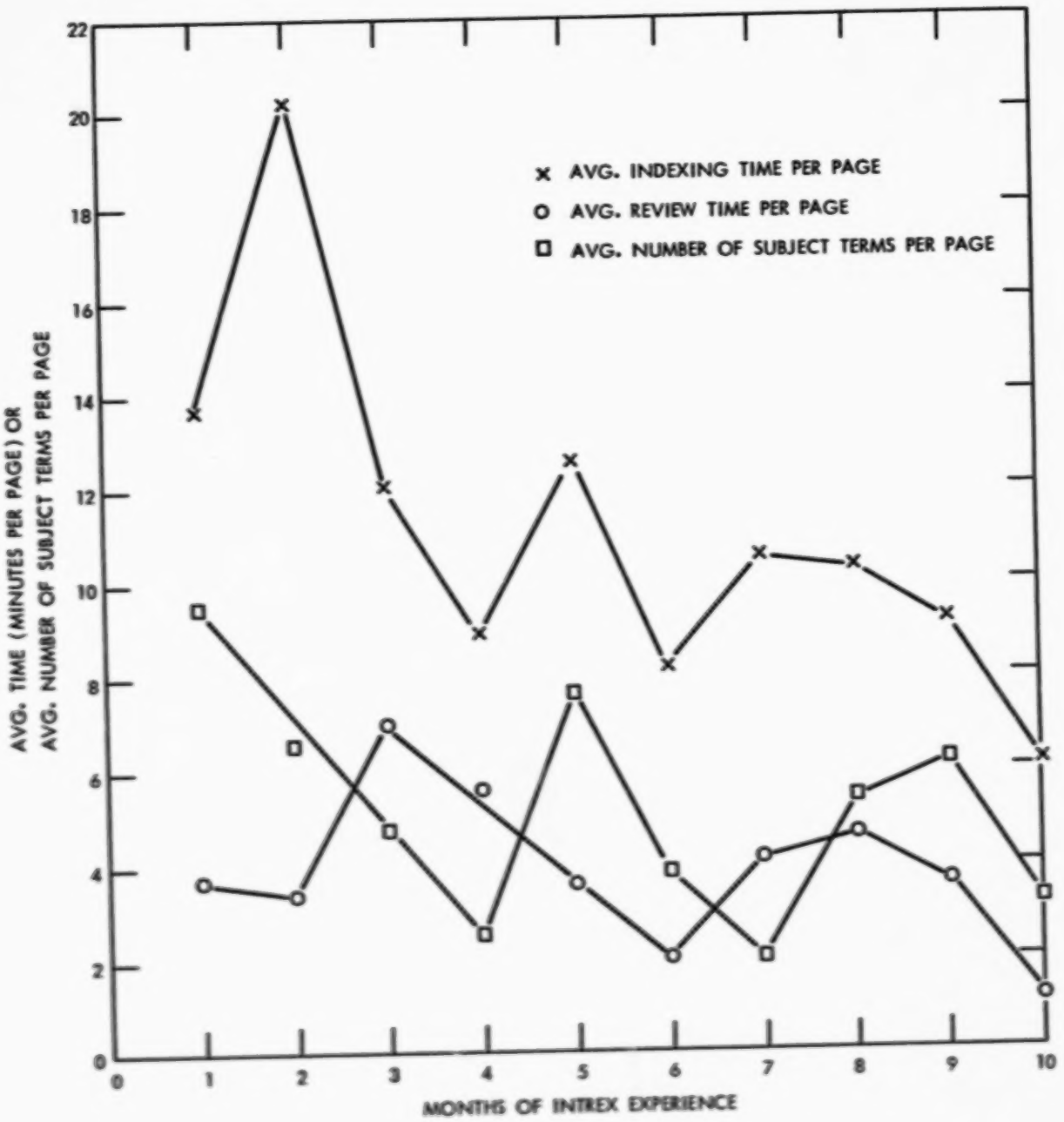


Fig. 5 Learning Curve and Other Time History Curves for Indexer 7

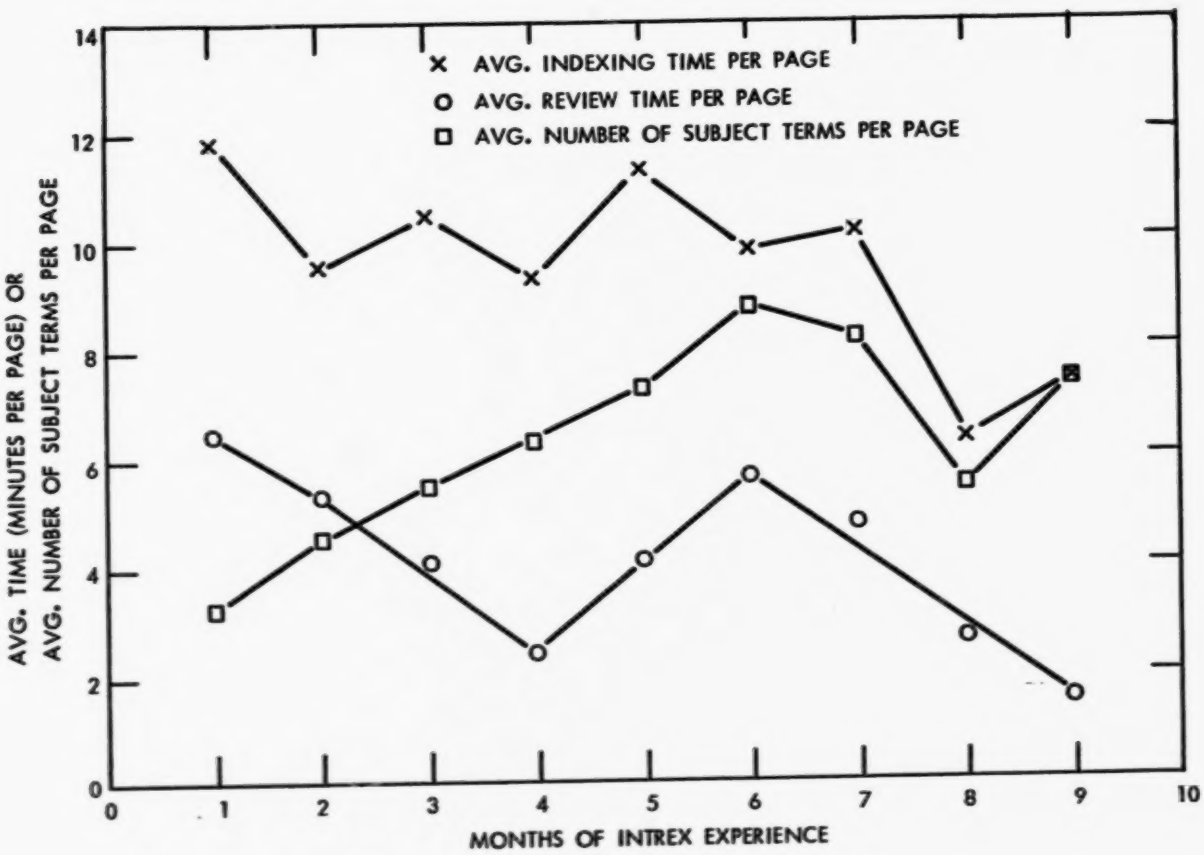


Fig. 6 Learning Curve and Other Time History Curves for Indexer 9

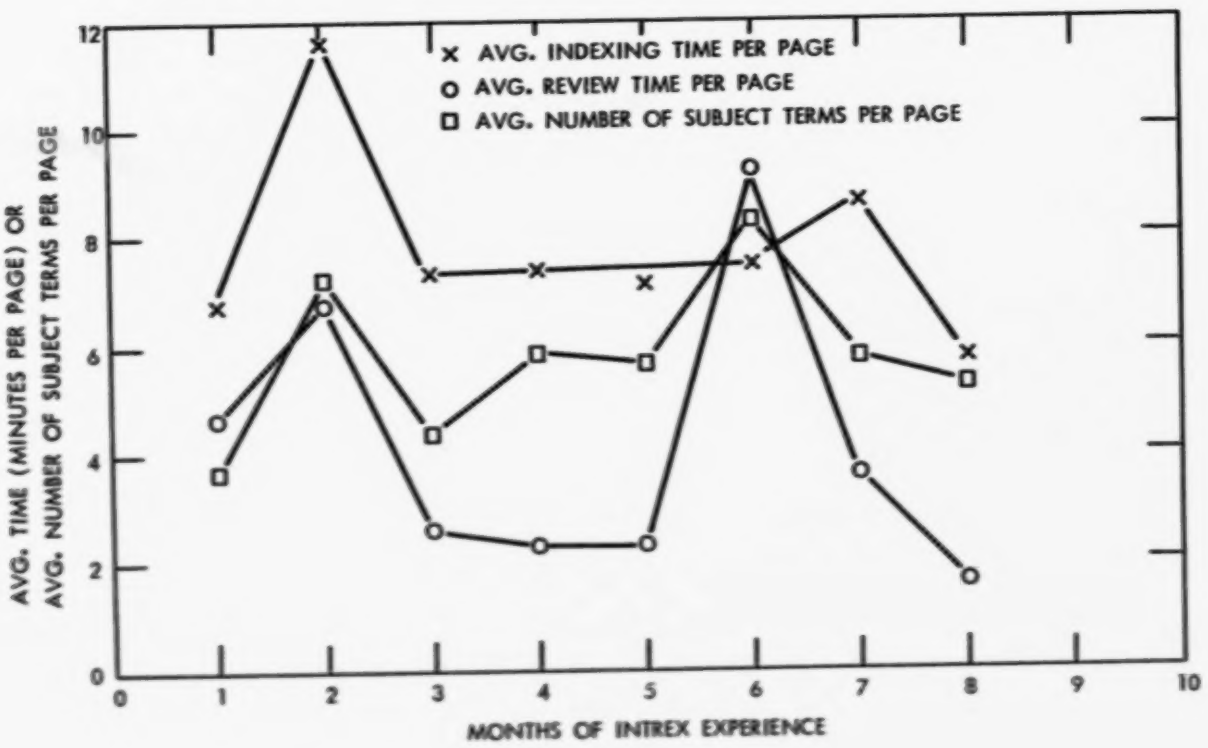


Fig. 7 Learning Curve and Other Time History Curves for Indexer 10

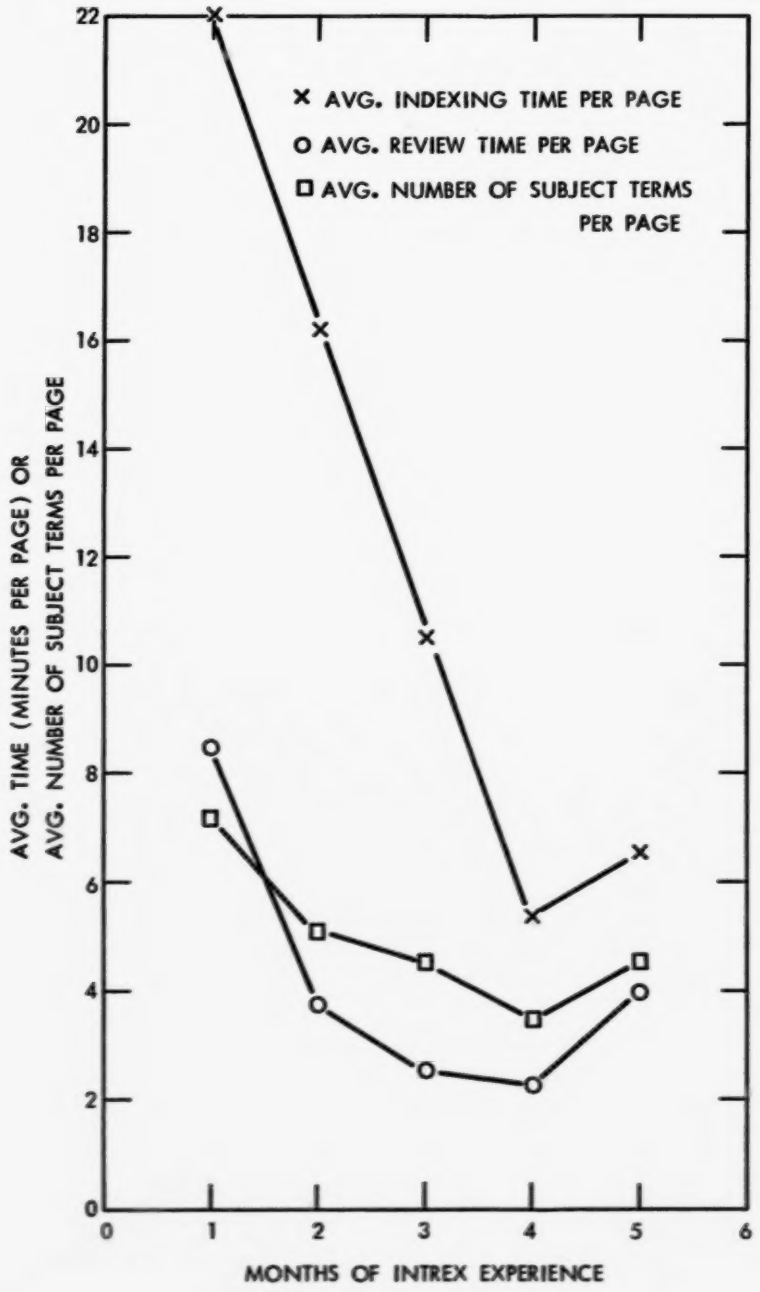


Fig. 8 Learning Curve and Other Time History Curves for Indexer 13

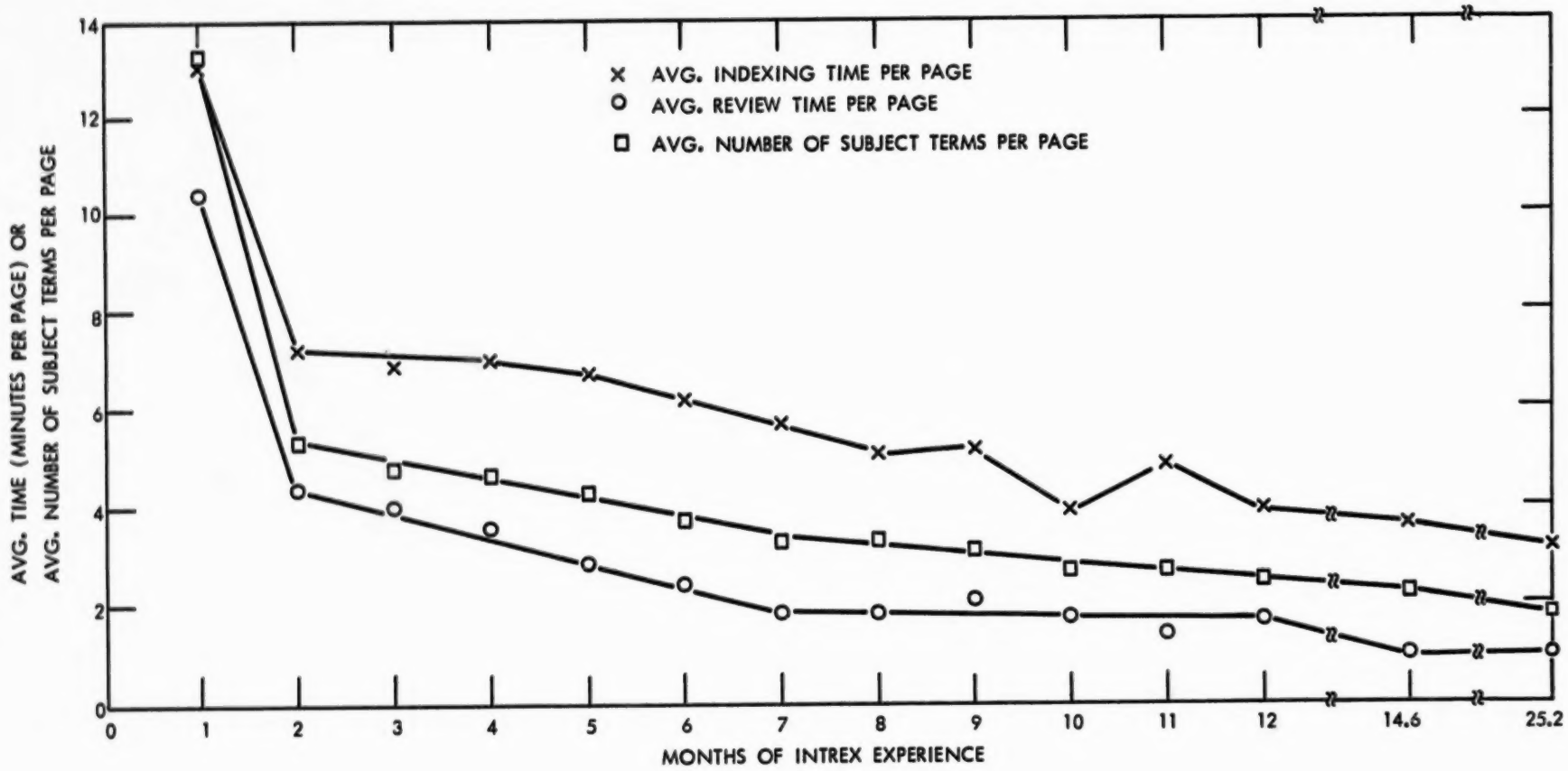


Fig. 9 Subject Indexing Time, Review Time, and Number of Subject Terms, per page, versus Document Length

Table 1

Learning Curve Data for All Indexers

Month Number	Number of Documents in Sample	Average Indexing Time min. per page	Standard Deviation
1	308	11.09	8.69
2	380	9.86	8.64
3	317	7.98	5.32
4	399	5.75	2.96
5	308	7.13	4.34

Table 2

Learning Curve Data for All Student Indexers

Month Number	Number of Documents in Sample	Average Indexing Time min. per page	Standard Deviation
1	166	10.40	9.90
2	155	8.66	7.78
3	76	5.76	2.84
4	38	5.76	2.38
5	12	6.43	4.01

Table 3

Learning Curve Data for All Librarian Indexers

Month Number	Number of Documents in Sample	Average Indexing Time min. per page	Standard Deviation
1	142	11.82	7.13
2	225	10.73	9.16
3	241	8.53	5.65
4	361	5.75	3.06
5	296	7.16	4.35
6	142	6.83	3.50
7	125	9.60	3.76
8	180	5.97	4.93
9	112	7.49	3.67
10	89	7.45	3.64
11	12	6.85	2.79
12	57	5.60	2.32
13	25	4.59	2.75
14	34	6.78	4.23

Table 4

Time History Data of the Mean per page Number of Subject Terms Assigned by All Indexers

Month Number	Number of Documents in Sample	Average Number of Subject Terms per page	Standard Deviation
1	308	5.34	3.48
2	380	4.97	3.02
3	317	4.24	2.57
4	399	4.39	2.25
5	308	4.82	3.27
6	142	5.12	3.02
7	125	6.25	2.73
8	180	4.28	2.70
9	112	5.57	3.81
10	89	4.16	2.12
11	12	5.95	4.26
12	57	3.69	2.24
13	25	2.69	1.72
14	34	4.01	2.68

Table 5

Learning Curve Data of Individual Indexers

Indexer Code Number	Month Number	Number of Documents in Sample	Average Indexing Time min. per page	Standard Deviation
1	1	16	7.25	3.40
	2	49	4.50	3.85
	3	16	4.44	1.50
	4	120	4.62	2.25
	5	86	4.21	2.16
	6	69	5.66	2.54
	8	42	3.79	2.46
	9	12	6.99	3.81
	10	12	7.58	2.54
	11	1	3.44	0.0
	12	32	4.84	2.06
	13	9	3.42	1.74
	14	3	4.22	1.17
2	1	6	3.94	1.52
	2	41	4.77	2.55
	3	60	4.86	3.31
	4	101	4.55	2.34
	5	95	6.18	2.45
	6	17	4.74	2.68
	7	5	3.75	0.75
	8	66	4.78	6.11
	9	33	5.68	2.66

Table 5 (Contd.)

Indexer Code Number	Month Number	Number of Documents in Sample	Average Indexing Time min. per page	Standard Deviation
2	10	53	8.10	4.07
	11	11	7.74	2.38
	12	25	6.57	2.28
	13	16	5.24	2.99
	14	31	7.03	4.34
7	1	49	14.71	5.98
	2	47	20.36	12.07
	3	44	12.10	4.86
	4	25	8.94	2.40
	5	31	12.56	4.18
	6	23	8.16	4.11
	7	50	10.52	3.41
	8	38	10.36	3.67
	9	40	9.15	4.43
	10	24	6.03	2.68
9	1	30	11.90	6.72
	2	35	9.56	4.50
	3	48	10.44	5.53
	4	35	9.35	3.86
	5	44	11.34	5.36
	6	26	9.84	3.31
	7	31	10.02	2.94
	8	16	6.33	1.90

Table 5 (Contd.)

Indexer Code Number	Month Number	Number of Documents in Sample	Average Indexing Time min. per page	Standard Deviation
9	9	27	7.39	2.02
10	1	30	6.80	2.67
	2	34	11.68	4.47
	3	36	7.44	3.70
	4	38	7.55	2.94
	5	33	7.13	2.97
	6	7	7.57	3.62
	7	39	8.69	4.23
	8	18	5.81	1.91
13	1	11	23.42	5.81
	2	19	16.23	7.52
	3	37	10.49	7.32
	[illegible]	[illegible]	[illegible]	[illegible]
	5	7	6.63	1.28
14	1	4	6.45	0.36
30	1	31	13.01	19.08
	2	72	10.75	10.02
	3	30	6.58	3.23
	4	19	4.34	1.33
32	1	13	9.05	5.63
	2	2	4.40	2.26
33	1	10	10.90	4.11
	2	34	9.92	5.07

Table 5 (Contd.)

Indexer Code Number	Month Number	Number of Documents in Sample	Average Indexing Time min. per page	Standard Deviation
33	3	24	5.95	1.29
34	1	11	16.92	5.81
35	1	16	9.50	4.07
	2	15	6.96	2.09
	3	5	6.70	2.60
	4	19	7.23	3.12
	5	12	6.34	4.17
36	1	14	8.29	3.45
37	1	14	5.38	2.72
	2	8	3.66	0.97
39	1	9	5.07	2.20
	2	24	4.27	2.25
	3	17	[illegible]	[illegible]
40	1	23	12.34	3.04
41	1	2	7.06	1.94
42	1	2	7.00	3.00
43	1	2	4.94	2.06
44	1	2	7.50	5.00
45	1	8	5.30	3.67
46	1	5	6.25	0.82

Table 6

Time History Data of the Mean Review Time
for Individual Indexers

Indexer Code Number	Month Number	Number of Documents in Sample	Average Review Time min. per page	Standard Deviation
1	2	5	1.21	0.41
	3	6	1.18	0.91
	4	120	2.15	1.71
	5	86	3.01	1.65
	6	69	7.31	5.78
	8	42	1.69	1.40
	9	12	3.32	1.70
	10	12	1.73	0.75
	11	1	0.78	0.0
	12	32	2.34	1.73
	13	9	2.17	2.32
	14	3	2.39	1.88
2	2	9	1.89	1.03
	3	41	1.62	1.09
	4	101	2.50	3.32
	5	95	3.19	1.99
	6	17	4.69	2.24
	7	5	1.80	0.68
	8	66	3.66	2.10
	9	33	3.79	2.48
	10	53	4.73	3.66
	11	11	5.49	7.56

Table 6 (Contd.)

Indexer Code Number	Month Number	Number of Documents in Sample	Average Review Time min. per page	Standard Deviation
2	12	25	2.83	1.43
	13	16	2.44	1.66
	14	31	3.01	2.10
7	1	49	3.79	2.95
	2	47	3.61	3.98
	3	44	6.99	5.77
	4	25	5.59	1.68
	5	31	3.64	2.78
	6	23	1.99	1.37
	7	50	4.07	3.26
	8	38	4.59	4.25
	9	40	3.68	3.70
9	1	30	6.51	4.37
	2	35	5.41	2.85
	3	48	4.10	3.11
	4	35	2.50	1.81
	5	44	4.11	4.27
	6	26	5.69	5.28
	7	31	4.84	3.54
	8	16	2.71	1.60
	9	27	2.90	1.97

Table 6 (Contd.)

Indexer Code Number	Month Number	Number of Documents in Sample	Average Review Time min. per page	Standard Deviation
10	1	30	4.65	3.01
	2	34	6.81	3.72
	3	36	2.70	1.99
	4	38	2.39	1.99
	5	33	2.35	1.42
	6	7	9.17	4.44
	7	39	3.60	2.90
	8	18	1.66	1.34
13	1	11	8.51	5.04
	2	19	3.88	2.56
	3	37	2.60	2.58
	4	42	2.34	2.04
	5	7	4.00	1.80
14	1	4	1.87	0.36
30	1	31	4.85	3.03
	2	72	4.18	2.44
	3	30	2.77	1.93
	4	19	2.85	2.05
32	1	15	4.76	3.12
33	1	10	2.94	1.25
	2	34	3.34	1.69
	3	24	3.98	1.16

Table 6 (Contd.)

Indexer Code Number	Month Number	Number of Documents in Sample	Average Review Time min. per page	Standard Deviation
34	1	11	5.04	1.37
35	1	16	3.27	1.41
	2	15	2.15	0.97
	3	5	4.40	1.32
	4	19	3.76	1.39
	5	12	5.66	1.53
36	1	14	4.99	2.44
37	1	14	4.36	5.84
	2	8	1.12	0.75
39	1	9	3.40	1.43
	2	24	2.26	1.21
	3	17	1.81	1.43
40	1	23	2.37	1.17
41	1	2	3.37	0.62
42	1	2	3.33	0.00
43	1	2	1.68	0.71
44	1	2	6.00	2.50
45	1	8	2.86	2.43
46	1	5	9.40	3.45

Table 7

Time History Data of the Mean per page Number of Subject Terms Assigned by Individual Indexers

Indexer Code Number	Month Number	Number of Documents in Sample	Average Number of Subject Terms per page	Standard Deviation
1	1	16	2.19	0.62
	2	49	2.20	0.97
	3	16	2.30	0.92
	4	120	3.73	1.62
	5	86	2.95	1.29
	6	69	3.98	1.80
	8	42	1.91	1.11
	9	12	3.89	2.36
	10	12	3.09	1.19
	11	1	2.33	0.0
	12	32	2.75	1.58
	13	9	2.31	2.10
	14	3	3.44	0.10
2	1	6	2.46	0.81
	2	41	3.23	1.55
	3	60	3.20	1.71
	4	101	3.88	1.55
	5	95	4.02	1.36
	6	17	4.05	1.33
	7	5	3.67	0.77
	8	66	4.57	3.02
	9	33	4.12	1.78

Table 7 (Contd.)

Indexer Code Number	Month Number	Number of Documents in Sample	Average Number of Subject Terms per page	Standard Deviation
2	10	53	4.84	2.23
	11	11	6.64	4.08
	12	25	4.89	2.38
	13	16	2.89	1.42
	14	31	4.07	2.81
7	1	49	9.55	4.47
	2	47	6.56	2.80
	3	44	4.64	2.33
	4	25	4.81	2.48
	5	31	7.63	4.39
	6	23	3.91	1.68
	7	50	5.64	2.12
	8	38	5.38	1.91
	9	40	6.03	1.48
	10	26	3.20	1.53
9	1	30	3.31	1.46
	2	35	4.79	2.12
	3	48	5.51	3.26
	4	35	6.14	2.56
	5	44	7.29	5.03
	6	26	8.79	3.57
	7	31	8.06	3.09
	8	16	5.47	2.04
	9	27	7.51	5.78

Table 7 (Contd.)

Indexer Code Number	Month Number	Number of Documents in Sample	Average Number of Subject Terms per page	Standard Deviation
10	1	30	3.71	1.52
	2	34	7.23	2.70
	3	36	4.37	1.95
	4	38	5.97	3.40
	5	33	5.72	2.64
	6	7	8.26	2.42
	7	39	5.91	2.54
	8	18	5.36	2.56
13	1	11	7.07	3.48
	2	19	5.08	1.94
	3	37	4.65	1.97
	4	42	3.56	1.34
	5	7	4.61	1.14
14	1	4	3.29	0.19
30	1	31	5.45	2.52
	2	72	5.92	3.86
	3	30	4.56	3.46
	4	19	5.13	2.82
32	1	13	8.08	3.71
	2	2	6.15	2.52
33	1	10	5.45	1.99
	2	34	6.44	2.99
	3	24	4.34	1.43

Table 7 (Contd.)

Indexer Code Number	Month Number	Number of Documents in Sample	Average Number of Subject Terms per page	Standard Deviation
34	1	11	4.88	2.04
35	1	16	6.82	2.66
	2	15	6.80	4.47
	3	5	5.53	1.88
	4	19	5.54	2.68
	5	12	5.55	3.38
36	1	14	4.05	1.14
37	1	14	3.88	1.68
	2	8	2.69	1.14
39	1	9	2.92	1.05
	2	24	2.13	1.07
	3	17	1.85	0.53
40	1	23	4.53	1.57
41	1	2	4.12	0.87
42	1	2	6.00	1.33
43	1	2	3.26	1.53
44	1	2	4.75	2.25
45	1	8	1.70	0.85
46	1	5	3.50	0.84

Table 8

Average Indexing Time, per page, versus Document Length, for All Experienced Indexers

Number of Pages	Number of Documents in Sample	Average Indexing Time min. per page	Standard Deviation
1	17	13.00	7.34
2	649	7.24	3.96
3	269	6.92	4.37
4	152	7.04	3.36
5	111	6.78	3.38
6	87	6.16	3.04
7	66	5.77	2.73
8	73	5.06	2.42
9	27	5.20	2.49
10	30	3.97	1.84
11	20	4.91	2.08
14.6 (average for sample of 13-19 pages)	32	3.59	1.32
25.2 (average for sample of 20-34 pages)	12	3.18	1.42
over 100	4	0.71	0.03

Table 9

Average Review Time, per page, versus Document Length, for All Experienced Indexers

Number of Pages	Number of Documents in Sample	Average Review Time min. per page	Standard Deviation
1	26	10.4[illegible]	7.13
2	910	4.45	3.71
3	445	3.99	3.14
4	237	3.59	2.59
5	169	2.98	2.12
6	125	2.39	1.63
7	97	1.95	1.16
8	97	1.81	1.30
9	58	2.03	1.49
10	56	1.78	1.33
11	37	1.45	1.02
14.6 (average for sample of 13-19 pages)	32	1.00	1.05
25.2 (average for sample of 20-34 pages)	12	0.91	0.64
over 100	4	0.13	0.02

Table 10

Average per page Number of Subject Terms Assigned versus Document Length, for All Experienced Indexers

Number of Pages	Number of Documents in Sample	Average Number of Subject Terms per page	Standard Deviation
1	17	13.35	8.86
2	649	5.36	2.65
3	269	4.85	2.91
4	152	4.76	2.39
5	111	4.33	1.99
6	87	3.76	2.05
7	66	3.32	1.71
8	73	3.41	1.63
9	27	3.02	1.58
10	30	2.68	1.63
11	20	2.62	0.89
12	16	2.46	0.92
14.6 (average for sample of 13-19 pages)	32	2.12	1.06
25.2 (average for sample of 20-34 pages)	12	1.80	0.89
over 100	4	0.42	0.03

Table 11

Average per page Indexing Time of Documents Grouped by Format

Format Category	Number of Documents in Sample	Average Indexing Time min. per page	Standard Deviation	Average Number of Pages
Textbook	4	0.71	0.30	282.75
Journal Article	605	6.83	4.00	6.34
Letters Journal Article	344	9.35	4.11	2.55
Conference Proceedings Article	597	5.10	2.56	2.86
Abstract	1	13.00	0.0	1.00
Bibliography	1	0.97	0.0	34.00
Reference Book	4	2.18	0.62	11.25

Table 12

Average per page Number of Assigned Subject Terms for Documents Grouped by Format

Format Category	Number of Documents in Sample	Average Number of Subject Terms per page	Standard Deviation	Average Number of Pages
Textbook	4	0.18	0.14	282.75
Journal Article	605	4.35	2.54	6.34
Letters Journal Article	344	6.58	3.64	2.55
Conference Proceedings Article	597	4.03	2.24	2.86
Abstract	1	9.00	0.0	1.00
Bibliography	1	1.62	0.0	34.00
Reference Book	4	2.09	1.30	11.25

Table 13

Average per page Indexing Time of Documents Grouped by the Author's Level of Approach

Level of Approach Category	Number of Documents in Sample	Average Indexing Time min. per page	Standard Deviation	Average Number of Pages
Professional in the author's subject area	290	5.82	3.93	3.29
Graduate Student	1235	6.91	3.85	5.31
College Senior	22	6.51	3.28	5.59

Table 14

Average per page Indexing Time of Documents Grouped by Subject Area

Subject Area Category	Number of Documents in Sample	Average Indexing Time min. per page	Standard Deviation	Average Number of Pages
Radio frequency, microwave, and optical spectroscopy of liquids and solids.	2375	7.98	6.09	5.21
High temperature metallurgy	116	5.93	3.09	2.32

Table 15

Average per page Indexing Time of Documents Grouped by Author's Purpose

Author's Purpose Category	Number of Documents in Sample	Average Indexing Time min. per page	Standard Deviation	Average Number of Pages
Report on theoretical research	355	6.79	4.23	5.33
Report on experimental research	504	6.74	3.62	3.17
Report on theoretical and experimental research	553	6.73	3.64	4.58
Report on a development or application	74	6.63	5.42	5.04
Critical review	12	6.88	5.65	7.00
Non-critical review	28	5.94	3.12	7.71
Critical comment	16	6.58	4.18	2.06
Non-critical comment	4	4.66	1.87	7.75
Textbook	3	0.57	0.22	303.00

APPENDIX B

Summary of Personal Data by Code Number

	Status; Date of Employment by Project Intrex; Field of Study:
1	Librarian; February 1967; English Literature (B.A.). Library Science (M.L.S.).
2	Librarian; February 1967; Fine Arts (B.A.). Library Science (B.A.).
7	Librarian; May 1967; History (B.A.). Library Science (M.L.S.).
9	Librarian; May 1967; Chemistry (B.S.). Library Science (M.L.S. plus Advanced Certificate).
10	Librarian; July 1967; Journalism (B.S.).
13	Librarian; November 1967; Secondary School Science Teaching (Certificate).
14	Student (Graduate); December 1967; Physics (B.S.).
30	Student (Senior); November 1967; Electrical Engineering.
32	Student (Senior); November 1967; Unknown.
33	Student (Junior); November 1967; Electrical Engineering.
34	Student (Junior); November 1967; Mechanical Engineering.
35	Student (Senior); November 1967; Physics.
36	Student (Junior); November 1967; Physics.
37	Student (Junior); November 1967; Electrical Engineering.
39	Student (Freshman); December 1967; Undecided.
40	Student (Senior); March 1968; Physics.
41	Student (Sophomore); February 1968; Aeronautical and Astronautical Engineering.
42	Student (Junior); March 1968; Chemical Engineering.
43	Student (Senior); March 1968; Chemistry.

	Status; Date of Employment by Project Intrex; Field of Study:
44	Student (Senior); March 1968; Biology.
45	Student (Junior); March 1968; Electrical Engineering.
46	Student (Junior); March 1968; Mathematics.

APPENDIX C

Sample Recording Form

Record Number	Subject Area	Number of Pages	Field 31 Code	Field 65 Code	Field 66 Code	Number of Terms in Field 73	Indexer	Date Indexed	Month Indexed Number	Subject Indexing Time	Special	Reviewer	Review Time
2312	1	007	BB	B	8	20	39	12-67	01	016	0	01	17
3730	1	005	BB	T	8	13	02	1-68	12	035	0	01	15
2820	2	010	CC	B	1	15	13	2-68	04	040	0	07	08
1472	1	023	DD	ED	9	47	07	7-68	02	075	1	10	23
Card Column Assignments:													
1-4	8	12-14	18-19	23-24	28	32-33	37-38		42-43	47-49	53	57-58	62-63

APPENDIX D

Listing of Files from which Data were Gathered

1 - 149

162

166

174

181

187

193

206

213

216

218

219

223

224

226 - 230

232 - 246

248 - 255

259

269

270

272

273

275

276

285

290 - 294

296 - 309

316 - 336

APPENDIX E

Computer Program

```
6  SUMSQ=0.
   M=0
   N=0
   SUM=0.
1  READ(5, 2) I, J
2  FORMAT (11X, I3, 47X, I2)*
   IF (I) 3, 4, 4
4  N=N+1
   M=M+1
   X=FLOAT(J)/FLOAT(I)
   SUMSQ=SUMSQ+X**2
   SUM=SUM+X
   GO TO 1
3  AVG=SUM/FLOAT(N)
   AGV=FLOAT(M)/FLOAT(N)
   VAR=SUMSQ/FLOAT(N)-AVG**2
   WRITE(6, 5) N, AVG, VAR, AGV
5  FORMAT (5H N = , I4, 5X, 6HAVG = , F11.6, 5X6HVAR = ,
   F11.6, F11.6)
   IF (J) 7, 6, 6
7  STOP
   END
```

*The independent parameter to be worked with (indexing time, review time, or number of subject terms) is specified by putting its data location (card columns) in the read statement.

APPENDIX F

A Note on Statistics

Individual variables, indexing time, review time, or number of subject terms per document page, are considered as independent, identically distributed, discrete random variables. Under these considerations, all data groupings are assumed to have a Gaussian probability distribution, although this assumption does not hold up well for less than about ten sample values. The mean (average) and standard deviation for all sample values in a data grouping were derived with the aid of computer processing and calculated as the sample mean and sample standard deviation for data fulfilling the above criteria. The formulations used were:

$$\text{Average} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\text{Standard Deviation} = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2\right)^{1/2}$$

where there are "n" documents represented in the data grouping and the "x_i's" are the sample values themselves, as for example, time to index divided by number of pages for a single record.

The standard deviation is a relative quantitative measure of the spread of the data about the mean. For a time Gaussian variable, some 68 percent of all sample points fall within the range of one standard deviation on either side of the mean. Many standard deviations in this study approach or exceed the associated mean values, telling us that our data does not conform to Gaussian theory by specifying a range approaching or passing through zero. Under these conditions, the standard deviation is still a valid measure of data spread; however, the data is somewhat skewed and percentages of data points lying within some range of the mean becomes uncalculable. That is,

relative magnitudes of deviations are still a qualitative measure of the spread of data about the mean.

Of particular interest in this study is the meaning of the extent of overlap of the standard deviations associated with different data groupings. This is largely a qualitative judgement to be made on the nature of the data groupings involved and the magnitude of the difference in mean values and extent of overlap.